Lead Scoring Case Study Summary

Stages in making a model

1. Data Gathering
2. Data Cleaning
3. Exploratory Data Analysis
4. Data Preparation
5. Model Building
6. Model Evaluation
7. Final recommendations

Data Gathering

Data set – Leads.csv

Data Dictionary – Leads Data Dictionary.xlsx

Data Cleaning

We checked for the percentage of NULL values in each column and straightaway **dropped columns having more than 70% NULL values**.

For the rest of the columns, we observed each column individually and either came up with a strategy to **impute its NULL values** or dropped the column.

Exploratory Data Analysis

We now had a data set free of NULL values. We started analyzing each column individually in Univariate analysis and made some observations.

Data Preparation

We then checked the columns having binary **(yes/No)** values and converted them to **1/0** and got **dummies for categorical** variables (and dropped the original categorical variables as well).

Model Building

We first create X and y data sets. **X** set contains all the **independent** variables and **y** set contains the dependent variable i.e. "**Converted**". We then split the data set into a train test split (**70**% - train data **30**% – test data) and got our training and testing sets.

We now used **Standard Scaler** (fit and transform) on certain numeric variables of our training data set to get their values in scale with all other columns. The training set contained of many columns as we had created dummies for all the categorical variables, so we eliminated features and selected the top ones using Recursive Feature Elimination (**RFE**).

A second measure was also taken and a **heatmap** was made of all the correlations between the data set, just to ensure if we are missing any high correlations. The heatmap revealed that there were no significant correlations that we were missing.

Now, we created the **final model** based on the **top 15** columns/features chosen earlier using RFE. Based on this model, we created a prediction and converted probability column and a **confusion matrix** of converted vs predicted.

Model Evaluation

The training model's **accuracy** came out to be **91.91%.** We then checked the **VIFs** of all the features which came out to be **less than 2**, so **no feature elimination** was required at this stage. We then calculated the **sensitivity**, **specificity**, and other metrics beyond accuracy.

We then plotted the **ROC curve** and understood that our models seemed to be good enough as it nearer to the top and left borders of the ROC curve and far away from the 45-degree diagonal. We then plotted the accuracy, sensitivity and specificity vs cut-off probability and found the **optimum cut-off point** to be **0.233.**

We then created two more columns **final_predicted** and **lead score** where we predicted for every person, based on the cut-off probability, and calculated the lead score based on the respective conversion probability we calculated earlier.

We then plotted Precision vs Recall trade-off and made predictions on the test set and found the model **accuracy** to be **81.5%.**

Final Recommendations

Based on the final model created, recommendations were given on how the conversion rate can be increased. Please refer to the PPT for the recommendations.